Report on Current Issues in Resilience

Nathan DeBardeleben, HPC-5

High-end computing (HEC) is requisite for solving our nation's most important scientific and engineering problems and has become increasingly vital to the mission of the national security community [1]. As the scale and complexity of HEC systems continue to grow, the impact of faults and failures will make it increasingly difficult to accomplish productive work using traditional means of fault-tolerance [2,3]. Further, the challenges of integrating large complex heterogeneous systems are increasing to the point where the stabilization period consumes a significant portion of the lifetime of those systems [4,5]. As a consequence of these two troubling trends, it will be necessary for the HEC community to identify innovative means for efficiently and affordably performing productive work on systems encountering frequent, persistent, and erratic errors –many of which will be undetectable by existing system-monitoring solutions. Resilience meets these daunting and ever-increasing challenges. To ensure the continued viability of the largest, most powerful, leading-edge computing systems will require standards-based solutions. These solutions must efficiently and dynamically guard and preserve information, computation, and data movement in the presence of faults and failures arising from complex system interactions and dependencies among platform hardware and software components, the system workload, and the physical environment.

The goal of HEC resilience is to enable effective and resource-efficient use of computing systems at extreme scale in the presence of system degradations and failures.

At the highest level, high-end computing is the process by which data is transformed into information through computation. Resilience facilitates this critical transformation process by accepting that the underlying hardware and software that comprises a system will be unreliable. In order to succeed, resilience assumes a new perspective in which uncertainty about the state of the system plays an important role in managing that system. Resources traditionally focused on maintaining a known and desirable system state are instead focused on end-to-end fidelity of data, computation, and data movement. Resilience is concerned with reliability of information in lieu of, or even at the expense of, reliability of the system. This novel approach to fault-tolerance is necessary in order to address the two-fold challenge of decreasing system reliability, because of increasing scale, and decreasing certainty about the operational state of the system due to increasing complexity. The resilience community proposes to address these challenges in five focused but overlapping thrust areas: (see Fig.) 1)theoretical foundations, 2) enabling infrastructure, 3) fault prediction and detection, 4) monitoring and control, and 5) end-to-end data integrity. To manage this prodigious scope, a successful program of resilience research will require coordinated, multidisciplinary undertakings in each of these thrust areas. This requirement forms the justification of a call for a national effort in resilience.

For more information contact Nathan DeBardeleben at ndebard@lanl.gov.

[1] Report of the *High End Computing Revitalization Task Force* (HECRTF), May (2004).
[2] B. Shroeder, G.A. Gibson, *J. Phys.: Proc. Sci. Disc. Adv. Comput. Prog. (SciDAC) Conf.* **78,** 2022 (2007); http://www.iop.org/EJ/abstract/1742-6596/78/1/012022.
[3] J.T. Daly, "Application Resilience for Truculent Systems," at *Fault Tolerance Workshop for Extreme Scale Computing (2009)*; http://www.teragridforum.org/mediawiki/images/8/80/Daly2009ws.pdf.
[4] "Roadrunner to Expand Hybrid Computing Applications," *ASC eNews Quarterly News Letter,* NA-ASC-500-09 Issue 10.
[5] B. Comes, B. Bland, "Testing & Integration," *Petascale Systems Integration into Large Scale Facilities Workshop* (2007); http://www.nersc.gov/projects/HPC-Integration/presentations/Breakout_3_Testing_Integration.ppt.

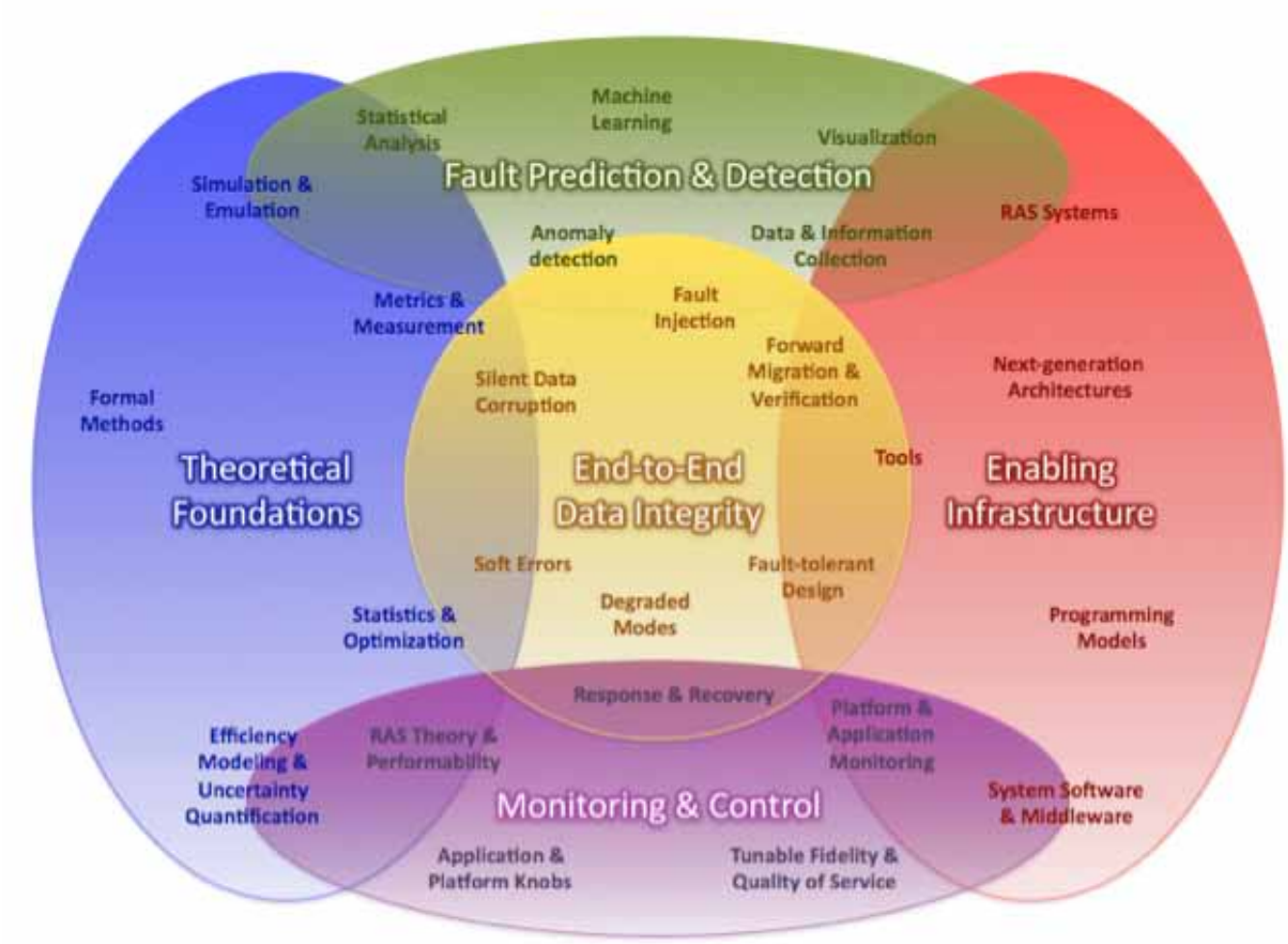

Fig. 1. Five focused but overlapping thrust areas.

For the full paper, see: http://institute.lanl.gov/resilience/docs/HECResilience_WhitePaper_Jan2010_final.pdf
For links to the talks presented at the workshop see: http://institute.lanl.gov/resilience/conferences/2009/